1. INTRODUCTION

1.1. Background

The recommendation system is the system that analyzes the data about of a product or user interaction with a product and then processes it to produce an output as a recommendation for users [1]. Many recommendation systems have been developed using collaborative filtering techniques and based on contents. Using these two techniques, user requirements do not need to be explicitly obtained from users. The system gets the needs of users by learning behavior based on the existing history (content-based) or learning other users who have the same taste as users (collaborative filtering). However, both techniques have cold start problems [2] - there are users who do not have product selection / history related to product as the others chosen - so the system will not be able to produce optimal results. Moreover, the user are rarely able to express their needs and also feel unsatisfied at the beginning of the first recommendation [3]..

Therefore, the Conversational recommender system (CRS) is proposed to overcome the cold start problem. The system uses other techniques besides the two previous techniques, namely Knowledge Based Recommended System. To obtain user needs, the system utilized repeated interactions between users and systems. It is inspired by conversations between a customer and a professional sales support, where the system will reportedly by providing questions / samples to the users. Some CRS studies [4, 5, 6, 7] have been conducted using many features in many domains. However, they do not cover all features and tend to use functional features. For example, in the smartphone domain, a user who wants a smartphone for daily needs such as reading documents but then he realizes that it cannot record videos.

Widyantoro and Baizal[4]propose a CRS framework for functional requirements, through Navigate with Request (NBA) and Navigation by Proposing (NBP) to guide users torefine their preferences during a conversation.The CRS is built using ontology as knowledge-based. The built-in ontology structure supports the mechanism of interaction betwen user and the system in CRS. It is able to explore the needs of users in terms of functional needs, and becomes a model in generating interaction. Interaction with CRS includes questions and answers, namely generating questions based on the feedbacks from the users, then assisting in making recommendation processes along with providing facilities for explanation of those recommendations. The knowledge base on CRS is based on a real world knowledge base service where the information on the topic such as product details and descriptions must always be up-to-date. If it is not updated it adapts to the growth of domain information, then it can be assumed that CRS as a system recommendation can provide recommendations that are not in accordance with actual knowledge. However, the information gathering for knowledge bases with certain specialities of study [4] is still processed manually. This maintenance process is time consuming and error prone [8].

While building a knowledge based from scratch isvery difficult, it is easier to reuse from the related knowledge[9].Therefore, automatic or semi-automatic process that can adapt to updates, finding and inserting information into the knowledgebase that matches a given ontology are needed. The task of inserting instance and relation into an existing ontology is called ontology population [8].For example, to maintain more than hundreds of smartphone products with various brands and updating manually the knowledge base whenever a new product is available can be very time consuming. Using ontology population that can update instances, the product details and descriptions, the managing a knowledge base will minimize the costs incurred for resources and the time needed to update the ontology.

There are several similar studies that have conducted population ontologies using website documents. Thomaz et. al [6] propose ontology population using Natural Language Processing, learning patterns to extract ontological class instances, as well as Confidence-weighted metrics and Similarity Measures. Yasmin, et al. [12] also proposed a method for conducting Ontology Population on website documents for the agriculturaldomain. In this research used the NLP method, then used General Architecture Text Engineering (GATE) and ANNIE Processing Resources to extract entities in the ontology.

Our approach differs from a formalized work in literature, because in this study conducted ontology population specifically on a site with documents that have tabular data format. From this point, our work is more similiar to the literature [7][13]. Shchekotykhin et al [13] have created an ALLRIGHT system that performs ontology population of tabular documents in the camera domain. Commonly instance information description in many cases in this domain is in tabular form. In general, the table is suitable for the human reader because it is very compact and has a precise form for displaying information. The extraction method using Natural Language Processing (NLP) does not work well to the tabular data. Typically tabular data does not contain complete sentences but rather the values of simple attributes that describe the features of an item. So from this form of tabular data, the NLP method does not have information that can be exploited. This research uses web crawlers to find factual information from various websites, and uses the X-Means Clustering Algorithm to get the desired product (in this case, 1 product can be represented by several webpage documents). Ovunc [7] has created OPPCAT, a framework that is used to perform population ontologies of tabular data on e-commerce stores and product catalysts. The function of the framework is to extract tabular data from PDF files that are converted into spreadsheets. In addition, this research investigates anomalies that often arise when extracting tabular documents.

Some interesting points from [13] are as follows. Their ontology population supports the search for documents using keywords by search engines. Moreover, the specification value of products stored in their ontology is in the form of a numerical value. But, Unlike thesse approaches, our work proposes the tabular web document approach for ontology population on CRS with the assumption that an instance is represented by specific website because the outcome of this work intent to be used as the part of CRS in [4]. Furthermore, in [4], the value of product specifications stored in the ontology is in the form of qualitative which must be acquired first by the domain of experts. For example, a products of smartphone has a HSP 42.2 / 5.76 Mbps network speed specification, LTE-A (3CA) cat18 1200/150 Mbps, then by domain expert, this specification must be specified in the quality range of low, medium or high. Then based on certain parameters the domain expert must also determine all the technologies that this product has into the categories of High End Technology, Medium High Technology, Medium Low Technology or included in Lowe End Technology. Our work propose to replace the role of the domain expert in that process stage. Role of the domain expert in acquitition knowledge process is, in [4] the domain expert determines the range of quality gradations from a specification and determines the value of the quality gradations of each product specification. Quality Gradation values represent the quality level of the specifications (e.g. High, Medium High, Medium Low, Low, etc). The range of quality gradations between specifications can be different. Suppose the range of quality gradations from the chipset is High, medium and low. While the range of quality gradations from the main camera is High, Medium High, Medium low and Low. This value will determine the relations of instances with other objects in ontology. The value of quality gradationshave a problem that it can change if there is a new technology that replaces it. New product models appear at any time in the market [10]. So a number of new technical features also appear to improve product functionality. For example a product A currently has a chipset specification with a high quality gradations value. In the future the level of quality specifications can not be the same because the company must always release new products with better quality to attract buyers.

Therefore, need a method that can group the specifications based on the similarity of data properties, so that each cluster can represent qualitatively from a specification. Clustering is a method for grouping data based on the similarity of data properties. K-Means is an algorithm of clustering that can divide data into K-groups. One of the advantages of this algorithm that match with gradation quality case is that an instance can move to another cluster when its centroid point isrecalculated [11]. However, traditional K-Means have a series of weaknesses. Therefore, to overcome this, in this study we will use the improved K-Means algorithm namely Bi-Layer K-Means Clustering Algorithm, which can overcome the problem that data in the same cluster are quite different, so that the accuracy of the data clustering can be promoted and can offer an efficiency of the traditional k-means algorithm [5]. In this study, Bi-Layer K-Means Clustering Algorithm has a cluster result that is better than other k-means algorithms (tri-level K-means, k-means *, k-menas ++, FGKM, FKMCUCD, and traditional K-means) .

The tabular data in this study has problem where a smartphone product has a different format between one specification to other then some of them use difrent unit. This can affect the process of extraction. Some of them can be directly used for clustering because the data type is numeric , and the measurement is same between one product and the others so that the value of these specifications is representative of the quality of the specifications. For example the specification for battery capacity has a value namely 3600 mAh. The number 3600 will be able to be used directly into the cluster. Because the other products also use same units in this specification. Examples of other specifications in same category are video recording quality, RAM, internal memory, screen resolution etc. However, there are also specifications which the value cannot visibly represent the quality and must use some preprocessing stages and need other reference to obtain the gradation quality. For example, a product has Exynos 7420 Octa for the Chipset specification value, the numerical value in this specification cannot represent the real quality and the units used in other products also vary. For example other products have chipsets specification like Apple a11 Bionic, Mt6765 Helio P35 etc the numeric data contained in the name is a representation of the type of the chipset model itself. For other cases, there are camera technology specifications for a smartphone with values namely 12 MP, f / 1.5, 26mm (wide), 1 / 2.55 ", 1.4µm, OIS, dual pixel PDAF. Other products write camera technology specifications like this 2mp, f/2.4, depth sensor. This value consists of several parameters that affect the quality of gradations such as Aperture, Sensor Size, Pixel Size and stabilization. In camera technology specification, betwen one product with another has the same unit (although there are some products that do not support several parameters), but the highest value for each parameter does not represent the high or low quality, then with the other parameters are same or the opposite (the lowest value are Better). Examples of other specifications in same category as chipset and camera technology are GPS, and screen protectors. Both of this specification are need information from other reference to obtain the gradation quality.

Table 1.1 Previous Work Related to Ontologi Population and Clusteirng

Method	Method	Proposed System
Ontology Population from Text Web Document	Natural Language Processing, Confidence-weighted metrix, Pattern learning, Similarity Measure	UMOPOW - An Unsupervised Method for Ontology Population [6]
	Natural Language Processing, General Architecture Text Engineering (GATE),	Ontology Population for Agricultural Domain – Durian Ontology

	ANNIE Processing Resources.	[12]
Ontology Population from Tabular Web Document	Web Crawler, Visual Table recognition, X-Means Clustering Algorithm	ALLRIGHT – Automatic Ontology Instantiation from tabular Web Documents [13]
	Solving anomalies in tabular data, tabula,	OPPCAT – semi automatic ontology population from tabular data [7]

1.2. Statement of the Problem

The formulation of the problem that will be the object of this research is as follows. Firstly, How to design a framework for Ontology Population in a Conversational Recommender System based on the Functional Requirements from tabular web documents? How to ensure that the resulting ontology still suitable according to Conversational Recommender System ontology requirements?

1.3. Objective

This work aims to design a framework for ontology population on Conversational Recommender Systems based on the Functional Requirements as in [4] from tabular web documents so its instantiation as ontology result can substitute manual ontology update on CRS. The framework include clustering process that employ the Bi-Layer K-Means Clustering Algorithm as part of knowledge acquisition. To reach the objective, it is necessary to analyse and check the Individual consistency of resulting ontology. Secondly, Analyze the resulting ontology still suitable according to CRS ontology requirements by checking the CRS Ontology Requirements.

1.4. Hypotheses

Based on research [13] design a framework for ontology population from tabular web documents so its instantiation as ontology result can substitute manual ontology update so we can ensure the resulting ontology remains consistent and suitable according to the Conversational Recommendation System Ontology Requirements.

1.5. Scope and Delimitation

The proposed approach and updated ontology are the output of this work, The ontology population using data sources with HTML semi-structure / table type. Information about smartphone products and the required descriptions are widely available on the website. Based on [21] and [4] www.gsmarena.com is often used as a reference to determine the specifications of smartphone so we use this site to collect the technical specification of the product. The product information that we use in this study is the product specification such as Chipset, battery, GPS, network speed, video record quality, internal memory, screen protector, loudspeaker, RAM, selfie camera resolution, main camera resolution, sceeen resolution, camera technology, screen technology, product performance and Screen size.To do the ontology population, this research uses the structure of existing ontology for Conversational Recommender System based on Functional Requirements [4]. Ontology used in this study uses the language format used by CRS ontology in research [4] that is using Indonesian with the aim to make it easier when testing by comparing the two ontologies.

1.6. Significance of the study

The outcome of this work can be one part of CRS Functional Requirement Based [4] as the maintenance process of its ontology whenever the information changing. The knowledge acquisition of ontology population can substitute knowledge acquisition by the domain expert. Furthermore, this work can be an example of ontology implementation for students, lecturers, and the others.

1.7. Contribution of this Study

This work contribution is in terms of enriching the domain of the ontology population research, i.e smartphone domain. The proposed approach attempts to add more insight about further preprocessing to convert quantitative data into qualitative data before the data is annotated on ontology. Moreover, such process substitute the role of domain expert in conducting knowledge acquisition.